(4)

HARVARD UNIVERSITY
DEPARTMENT OF STATISTICS

FROM SPECIES PROBLEM TO A GENERAL COVERAGE PROBLEM VIA A NEW INTERPRETATION

BY

SHAW-HWA LO

HARVARD UNIVERSITY

Technical Report No. ONR-C-3

March 1989

Prepared under ONR Grant N00014-86-K-0246
Supported in part by NSF Grant DMS-88-17204

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION Unclassified	1b RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY	3 DISTRIBUTION / AVAILABILITY OF REPORT Unlimited
2b. DECLASSIFICATION / DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S) TR NO. ONR-C-3	5 MONITORING ORGANIZATION REPORT NUMBER(S)
6a. NAME OF PERFORMING ORGANIZATION Department of Statistics Harvard University 6b. OFFICE SYMBOL (If applicable)	7a NAME OF MONITORING ORGANIZATION
6c. ADDRESS (City, State, and ZIP Code) Department of Statistics, SC713 Harvard University Cambridge, MA 02138	7b. ADDRESS (City, State, and ZIP Code)
8a. NAME OF FUNDING / SPONSORING ORGANIZATION ONR 8b. OFFICE SYMBOL (If applicable) Code 1111	9 PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER N0001486-K-0246
8c. ADDRESS (City, State, and ZIP Code) Office of Naval Research Arlington, VA 22217-5000	10 SOURCE OF FUNDING NUMBERS PROGRAM ELEMENT NO. / PROJECT NO. / TASK NO. / WORK UNIT ACCESSION NO

11. TITLE (Include Security Classification)
From Species Problem to a General Coverage Problem Via a New Interpretation

12. PERSONAL AUTHOR(S)
Shaw-Hwa Lo

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Year, Month, Day)	15. PAGE COUNT
Technical Report	FROM _____ TO _____	March, 1989	31

16. SUPPLEMENTARY NOTATION

17 COSATI CODES			18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
FIELD	GROUP	SUB-GROUP	

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

SEE REVERSE SIDE

20 DISTRIBUTION / AVAILABILITY OF ABSTRACT ☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT ☐ DTIC USERS	21 ABSTRACT SECURITY CLASSIFICATION Unclassified
22a NAME OF RESPONSIBLE INDIVIDUAL Herman Chernoff	22b TELEPHONE (Include Area Code) 617-495-5462 / 22c. OFFICE SYMBOL

DD FORM 1473, 84 MAR 83 APR edition may be used until exhausted
All other editions are obsolete

SECURITY CLASSIFICATION OF THIS PAGE
UNCLASSIFIED

ABSTRACT

A new interpretation is given, which provides another way of understanding the structure of the species problem and sheds light on the properties of a general coverage problem. As an illustrative example, the popular Turing-Good-Robbins estimator is "shown" to be a natural choice from this interpretation in the species problem. We set up a general framework of various coverage problems in this paper. The new interpretation is applied to this general situation which leads to many interesting applications in addition to the species problem. The coverage problems considered in this paper include the species problem, the problem of estimating the volume of a convex set, and the missile-coverage problem. It is pointed out that the genral estimators derived from this new interpretation usually estimate the probabilistic phenomenon involving only "$n-1$" observations which may not be appropriate. A general modified procedure is thus suggested to improve the current estimators. To justify the interpretation theoretically, we present some limit theorems in terms of species problem, even though the results are expected to hold more generally.

Summary. A new interpretation is given, which provides another way of understanding the structure of the species problem and sheds light on the properties of a general coverage problem. As an illustrative example, the popular Turing-Good-Robbins estimator is "shown" to be a natural choice from this interpretation in the species problem. We set up a general framework of various coverage problems in this paper. The new interpretation is applied to this general situation which leads to many interesting applications in addition to the species problem. The coverage problems considered in this paper include the species problem, the problem of estimating the volume of a convex set, and the missile-coverage problem. It is pointed out that the genral estimators derived from this new interpretation usually estimate the probabilistic phenomenon involving only "$n-1$" observations which may not be appropriate. A general modified procedure is thus suggested to improve the current estimators. To justify the interpretation theoretically, we present some limit theorems in terms of species problem, even though the results are expected to hold more generally.

1. Introduction

The problem of estimating the total probability of unseen species goes back to A.M. Turing according to Good (1953). To describe the problem comprehensively, we use the notation of Robbins (1956, 1968). Let $\{e_1, e_2, e_3 \ldots\}$ be the possible distinct species with probabilities $p_1, p_2, \ldots$, being selected in a single experiment. In n independent trials suppose that n_r species appear r times, r=1,2,..., and $\sum_{r=1}^{\infty} r n_r = n$. We also use n_0 to denote the number of species which are not present in the sample. It is clear that $n_1, n_2, \ldots$, are observable, but n_0 is not. In fact n_0 is infinite if there are infinitely many species. Let $\{X_i = j\}$ if and only if the i^{th} trial results in outcome e_j.

For $r \geq 0$, let $\varphi_j(r;n)$=1 if the number of $\{X_i = j\}$ is r and 0 otherwise. In particular, the sum of the probabilities p_j for those species which are not observed is

$$C_0 = \sum_{j=1}^{\infty} p_j \varphi_j(0;n) \,. \tag{1.1}$$

More generally, the sum of the probabilities of all species that are each represented $r(r \geq 0)$ times in the sample is

$$C_r = \sum_{j=1}^{\infty} p_j \varphi_j(r;n) \,. \tag{1.2}$$

To estimate C_r, Turing (see Good (1953)) suggested the formulas:

$$\frac{(r+1)n_{r+1}}{n} \qquad \text{for } r \geq 0\,. \tag{1.3}$$

Using a uniform prior, Good (1953) gave a derivation of these estimators from a Bayesian point of view. Since then several other interpretations of these estimators have appeared in the literature. These include Good (1953), Robbins (1956, 1968), and Diaconis and Stein (1983) among others. Various justifications of this type of estimator have been given. It should be noted that Robbins (1968) constructed an "unbiased" estimator for C_0 which is very similar to (1.3). However, Robbins' estimator is justified through the device of adding an additional trial to the original n observations. Here an estimator is called "unbiased" for estimating a random variable if E(estimate) = E(random variable). The problem continues to attract the attention of many researchers. To name a few: Starr (1979), Clayton and Frees (1987), Estey (1986), Bickel and Yahav (1985), and Cohen and Sackrowitz (1988). Most works concern the properties of the estimators of type (1.3); either from asymptotic or decision theory points of view. As an important application, the species problem is currently of great interest to researchers in automated speech identification (Bahl et al (1983), Jelinek (1976), and Katz (1987) among others).

My object is to introduce another interpretation of these estimators which leads to interesting applications other than the species problem. Later in this section we shall outline my approach using the species problem as an illustrative example. As a consequence it will become quite clear why the estimators of type (1.3) are "natural choices" in the species problem.

In Section 2 a framework for a general coverage problem is introduced. Some general estimates and their properties are derived using my interpretation. It is pointed out that the general estimates (including (1.3) in the species problem) derived from the interpretation are usually "biased" slightly upward. A general modified procedure is suggested to reduce the biases. The success of this procedure depends heavily upon the nature of the underlying problems. Although the biases are relatively small for many applications, their reduction seems to be interesting from a theoretical point of view.

Section 3 consists of three subsections, 3.1-3.3, which display three special examples as direct applications of the general framework established in Section 2. It seems to this author that the range of potentially useful applications is broader than presented here. The

first example is a further discussion of the species problem. The second example concerns the problem of estimating the volume of an arbitrary convex figure in Euclidean space. The connection between the interpretation and the problem of estimating the volume of a convex polyhedron was pointed out to me by Diaconis in a conversation. Some new results related to this problem on the plane are given, and the structure of the problem on higher dimensions is sketched heuristically. The last example deals with a missile-coverage problem:

> "n missiles are delivered and landing at a certain target area which is usually larger than the 'effective area' caused by the explosion of a single missile. The typical questions we are interested in are: (1) if the $(n+1)^{th}$ missile is fired, what is the chance that this additional missile would involve area which was not covered previously? (2) How large is the newly covered area? (3) How many more missiles are needed to cover 90% of the target area?"

We shall provide most of the answers to these questions in Section 3.3.

Section 4 is rather technical, where we shall give some limit theorems in terms of the species problem. In order to present the idea simply and clearly, we have chosen to treat special cases, even though the results are expected to hold more generally.

The main purpose of this paper is to set up a framework including various coverage problems so that the relevant parameters can be estimated by estimators which are obvious choices through the interpretations. Now we shall use the species problem as an example to give the flavor of the interpretation.

Suppose we are interested in the probability C_r in the species problem. Let X_{n+1} denote the additional observation. The random probability C_r is identical to the following conditional probability

$$P\{X_{n+1} \in S_n(r) | X_1, X_2, \ldots, X_n\} \tag{1.4}$$

where $S_n(r) = \{ j; \varphi_j(r;n) = 1\}$. Based on n observations, it is natural to estimate

$$P\{X_j \in S_{n-1,j}(r) | A_{n,j}\} \text{ for all } 1 \le j \le n \tag{1.4'}$$

$$\text{by} \quad I_{S_{n-1,j}(r)}(X_j)\,,$$

where

$$A_{n,j} = \bigcup_{i \neq j}^{n} \{X_i\} \ , \quad S_{n-1,j}(r) = \{i; \varphi_{ij}(r; n-1) = 1\}$$

and $\varphi_{ij}(r; n-1) = 1$ if and only if i appear exactly r times in $A_{n,j}$.

With continuity property, it is expected that (1.4) and (1.4′) are close to each other. (A general discussion of this "closeness" is given in Section 2.) It is then natural to estimate

(1.5) $$\frac{1}{n} \sum_{j=1}^{n} P\{X_j \in S_{n-1,j}(r) | A_{n,j}\} \quad \text{by}$$

(1.6) $$\frac{1}{n} \sum_{j=1}^{n} I_{S_{n-1,j}(r)}(X_j) \ .$$

Since

(1.7) $$I_{S_{n-1,j}(r)}(X_j) = 1 \iff X_j \in S_n(r+1) \ ,$$

the estimate (1.6) can thus be rewritten as

(1.8) $$\frac{(r+1)n_{r+1}}{n} \ ,$$

which is exactly the formula suggested by Turing and studied by Good (1953, 1956).

By taking expectation, we obtain

(1.9) $$\mathrm{E}(P\{X_{n+1} \in S_n(r) | X_1, X_2, \ldots, X_n\}) = P\{X_{n+1} \in S_n(r)\}$$

(1.10) $$\mathrm{E}(P\{X_j \in S_{n-1,j}(r) | A_{n,j}\}) = P\{X_n \in S_{n-1}(r)\}, \text{and}$$

(1.11) $$\mathrm{E}(\frac{1}{n} \sum_{j=1}^{n} I_{S_{n-1,j}(r)}(X_j)) = \mathrm{E}\,(\frac{(r+1)n_{r+1}}{n}) = P\{X_n \in S_{n-1}(r)\} \ .$$

Therefore, $\frac{(r+1)n_{r+1}}{n}$ is an "unbiased" estimate of $P\{X_n \in S_{n-1}(r) | X_1, X_2, \ldots, X_{n-1}\}$ in the sense that both random quantities have the same expectation. This is contrasted with the Robbins' arguments (1967) where "unbiasedness" was proved in the case r=0 through direct calculations. Here, the "unbiasedness" is shown more generally with no calculation.

We saw that the estimator is the average of naive estimators based on samples of size $n-1$, and it estimates $P(X_n \in S_{n-1}(r))$, a probabilistic statement based on "$n-1$" observations. As an estimator of (1.4), (1.8) is biased. This bias is slight because (1.4) changes little as n increases. In Section 3 we shall improve this estimator to reduce the bias which could be substantial in other problems for which this approach applies.

The key idea of this approach is to create requred information by temporarily deleting one observation from the sample one at a time, and the required information is obtained by comparing the deleted observation with the remaining $n-1$ observations. The final estimate is obtained by taking the average over these n steps, and it is no surprise that the final estimator really estimates the probabilistic phenomenon involving $n-1$ observations. Even though the idea behind the procedure is simple, it can be generalized to a fairly general model which is the subject of the next section.

2. A General Coverage Problem

In this section we shall discuss a general coverage problem in which a random sample $X_1, \ldots, X_n$ of size n is observed from a certain probability space (Ω, F, P). Let Ω denote a collection of certain subsets of a fixed set Δ in $\Re^k$, $k \geq 1$, whereas F and P are an appropriate δ-field and a probability measure defined on F.

Typical sample outcomes of $X_1, \ldots, X_n$ are n subsets of Δ. Consider all possible finite intersections among $\{X_i\}_{i=1}^n$ and Δ, it is easy to check that these intersections result in a finite partition $\pi_n = \{\wedge_i\}_{i=1}^{2^n}$ of Δ with 2^n disjoint subsets of Δ. Let g be a well-defined function from Ω to R^k. Some of the problems we wish to solve are the following:
Given a specified subset $S_n = S(X_1, X_2, \ldots, X_n; P)$ of Δ, possibly depending on both $\mathbf{X}_n = (X_1, X_2, \ldots, X_n)$ and P, estimate

(i) the probability that $g(X_{n+1}) \in S_n$ given S_n. Furthermore, if all elements in Ω are Lebesque-measureable, we are interested in estimating

(ii) the expected volume of $S_n \cap X_{n+1}$ given S_n, and

(iii) the expected volume of $S_{n+1} \cap S_n$ given S_n if additional sample X_{n+1} is made.

We assume throughout this section that S_n is defined for every $n \geq 1$. The key idea can best be described as a one-step "backward" procedure as follows. Let X_j be randomly removed from the sample $\{X_1, X_2, \ldots, X_n\}$, and let $A_{n,j}$ denote the j^{th} removed sample, i.e, $A_{n,j} = \bigcup_{i \neq j}^{n} \{X_i\}$. Let $S_{n-1,j} = S(A_{n,j}; P)$ be the specified subset of Δ based on the sample

$A_{n,j}$ of size $n-1$. We further define an indicator function

$$I_{S_{n-1,j}}g(X_j) = \begin{cases} 1 & \text{if } g(X_j) \in S_{n-1,j} \\ 0 & \text{otherwise.} \end{cases}$$

As pointed out in the previous section, our procedure will lead to some estimators which estimate the probabilistic statements involving "$n-1$" observations. For this reason, we shall call them "$(n-1)$-estimators" hereafter.

Instead of estimating the probability $P(g(X_{n+1}) \in S_n|S_n)$ in (i), the "$(n-1)$-estimator" estimates $P(g(X_n) \in S_{n-1}|S_{n-1})$. The construction can be described as follows: (i') In order to estimate $P(g(X_n) \in S_{n-1}|S_{n-1})$, note that the probability that $g(X_j) \in S_{n-1,j}$ can be estimated by $I_{S_{n-1,j}}[g(X_j)]$ empirically. In fact, this estimator is "unbiased" in the sense that

$$\mathrm{E}(I_{S_{n-1,j}}[g(X_j)]) = P\{g(X_n) \in S_{n-1}\} = \mathrm{E}(P(g(X_n) \in S_{n-1}|S_{n-1}))\ .$$

Since X_j is randomly removed from the sample, a final estimator ($(n-1)$-estimator) is thus

$$\frac{1}{n}\sum_{j=1}^{n} I_{S_{n-1,j}}[g(X_j)]\ ,$$

which is also "unbiased."

Likewise, instead of estimating (ii) we estimate

(ii') $$\mathrm{E}(\mathrm{vol}\ [S_{n-1} \cap X_n]|S_{n-1})\ .$$

Consider the estimator vol $[S_{n-1,j} \cap X_j]\ \ \forall\ 1 \le j \le n$. It is clear that

$$\begin{aligned} &\mathrm{E}\ (\mathrm{vol}[S_{n-1,j} \cap X_j]) \\ &= \mathrm{E}\ (\mathrm{vol}[S_{n-1} \cap X_n]) \\ &= \mathrm{E}\ (\mathrm{E}\ (\mathrm{vol}[S_{n-1} \cap X_n]|S_{n-1}) \\ &\forall\ 1 \le j \le n\ , \end{aligned}$$

the $(n-1)$-estimator is thus

$$\frac{1}{n}\sum_{j=1}^{n} \mathrm{vol}[S_{n-1,j} \cap X_j]$$

and is also "unbiased."

For estimating

(iii') $$\mathrm{E}\,(\mathrm{vol}[S_n \cap S_{n-1}]/S_{n-1})\ ,$$

we consider the estimator vol $[S_{n-1,j} \cap S_n]$ $\forall\ 1 \le j \le n$. Again, it is easy to see

$$\begin{aligned}\mathrm{E}\,(\mathrm{vol}[S_{n-1,j} \cap S_n]) &= \mathrm{E}\,(\mathrm{vol}[S_n \cap S_{n-1}]) \\ &= \mathrm{E}\,(\mathrm{E}(\mathrm{vol}[S_n \cap S_{n-1}]|S_{n-1}))\end{aligned}$$

$$\forall\ 1 \le j \le n\ .$$

The final $(n-1)$-estimator is

$$\frac{1}{n}\sum_{j=1}^{n} \mathrm{vol}[S_{n-1,j} \cap S_n]\ .$$

Remark. The assumptions made above about the sampling plan can be further relaxed. In fact, one can check that the only assumption we need (to guarantee the conclusion) is $L(X_1,\ldots,X_n|P)$=$L(X_{\pi 1}, X_{\pi 2},\ldots,X_{\pi n}|P)$ for any permutation π on $\{1,2,\ldots,n\}$ for every n. In particular, if $(X_1,\ldots,X_n)$ are exchangeable random elements, all the conclusions discussed above still hold.

If $S_n = S(X_1, X_2,\ldots,X_n; P) = S(P)$ does not depend on $X_n = (X_1, X_2,\ldots,X_n)$, it is easy to check that our interpretation will lead to an estimator which is the well known estimator obtained by the empirical measure.

As estimators of (i), (ii), and (iii), these $(n-1)$-estimators are all "biased." In many applications the biases are slight because (i), (ii), and (iii) changed little as n increases. We shall refer to this property as continuity property. However, in our general framework, this property is not automatically guaranteed. As a result, just how well these $(n-1)$-estimators estimate (i), (ii), and (iii) depends upon the forms of S_n and S_{n-1}. The following proposition tells us that the sucess of using $(n-1)$-estimators to estimate (i), (ii), and (iii) depends on the "closeness" of S_{n-1} to S_n.

Proposition 2.1. Assuming that X is randomly chosen from (Ω, F, P) and is independent of $S_{n-1}(X_1, X_2,\ldots,X_{n-1}; P)$, $S_n(X_1, X_2,\ldots,X_n; P)$ and $S_{n+1}(X_1, X_2,\ldots,X_n, X_{n+1}; P)$. Let g be a measurable function from (Ω, F, P) to $\Re^k$ such that $g(w) \in w$ for all $w \in \Omega$. We further assume $\mathrm{E}\ \mathrm{vol}(X)^2 < \infty$.

If

$$P\{[X \cap \{S_n \cup S_{n-1} \backslash S_n \cap S_{n-1}\}] \neq \phi\} = \delta_n \geq 0 \text{ for all } n \geq 1 ,$$

then

$$(1)\ |P\{g(X_{n+1}) \in S_n\} - P\{g(X_n) \in S_{n-1}\}| \leq \delta_n \text{ and}$$

$$(2)\ \mathrm{E}\,(\mathrm{vol}[S_n \cap X_{n+1}]) - \mathrm{E}\,(\mathrm{vol}[S_{n-1} \cap X_n]) = 0(\delta_n^{\frac{1}{2}}) .$$

If we further assume vol $(\Delta) < \infty$, then (2) becomes

$$(2')\ \mathrm{E}\,(\mathrm{vol}[S_n \cap X_{n+1}]) - \mathrm{E}\,(\mathrm{vol}[S_{n-1} \cap X_n]) = 0(\delta_n) .$$

Proof of (1)

It suffices to show

$$|P\{g(X) \in S_n\} - P\{g(X) \in S_{n-1}\}| \leq \delta_n .$$

Since

$$P\{g(X) \in S_n\} - P\{g(X) \in S_{n-1}\}$$
$$= P\{g(X) \in S_n \backslash S_{n-1}\} - P\{g(X) \in S_{n-1} \backslash S_n\}$$

it follows from assumptions that both terms above are smaller than

$$P\{[X \cap \{[S_n \cup S_{n-1}] \backslash [S_n \cap S_{n-1}]\}] \neq \phi\} = \delta_n ,$$

and the proof of (1) follows immediately.

Proof of (2) One can write

$$|\mathrm{E}\,(\mathrm{vol}[S_n \cap X_{n+1}]) - \mathrm{E}\,(\mathrm{vol}[S_{n-1} \cap X_n])|$$

$$= |\mathrm{E}\,(\mathrm{vol}[S_n \cap X] - \mathrm{vol}[S_{n-1} \cap X])|$$

$$= |\mathrm{E}\,(\mathrm{vol}[S_n \cap X) \backslash (S_{n-1} \cap X)] - \mathrm{E}\,(\mathrm{vol}[(S_{n-1} \cap X) \backslash (S_n \cap X)])|$$

Both of the above terms are clearly bounded by

$$\mathrm{E}\,(\mathrm{vol}(X \cap [S_n \cup S_{n-1}] \backslash [S_n \cap S_{n-1}]))$$
$$\leq \mathrm{E}\,[\mathrm{vol}(X) \cdot I_{[S_n \cup S_{n-1}] \backslash [S_n \cap S_{n-1}]}(X)]$$
$$\leq \mathrm{E}\,[\mathrm{vol}(X)^2]^{\frac{1}{2}} \cdot \delta_n^{\frac{1}{2}} = 0(\delta_n^{\frac{1}{2}}) .$$

This completes the proof of (2). *If* vol $(\Delta) < \infty$, *then since* vol $(X) \leq$ vol(Δ) w.p.1., it follows that

$$\mathrm{E}\,[\mathrm{vol}(X) I_{[S_n \cup S_{n-1}] \setminus [S_n \cap S_{n-1}]}(X)] \leq \ \mathrm{vol}(\Delta) \cdot \delta_n = 0(\delta_n)\ ,$$

which completes the proof of (2').

The "Biases" of $(n-1)$-estimates

As we have shown, in (i), (ii), (iii), the proposed $(n-1)$-estimates are "unbiased" in estimating the probabilistic statements (involved only $n-1$ observations), which are different from those based on n observations. In other words, there would be some biases if we use these $(n-1)$-estimates.

To calculate the biases, we pretend the additional observation, X_{n+1} is taken. The (n)-estimates obtianed by applying (i), (ii), and (iii) to this $n+1$ observation should be "unbiased." Therefore, the biases of $(n-1)$-estimates can be evaluated by comparing these $(n-1)$-estimates with (n)-estimates. For example, as in (i), the $(n-1)$-estimate is

$$\frac{1}{n} \sum_{j=1}^{n} I_{S_{n-1,j}}[g(X_j)]\ ,$$

and the (n)-estimate is

$$\frac{1}{n+1} \sum_{j=1}^{n+1} I_{S_{n,j}}[g(X_j)]\ .$$

The "bias" of $(n-1)$-estimate is thus

$$\mathrm{E}\,\{\frac{1}{n} \sum_{j=1}^{n} I_{S_{n-1,j}}[g(X_j)] - \frac{1}{n+1} \sum_{j=1}^{n+1} I_{S_{n,j}}[g(X_j)]\}\ . \tag{2.1}$$

where

$$S_{n,j} = S(A_{n+1,j}, p), \text{ and } A_{n+1,j} = \bigcup_{i \neq j}^{n+1} \{X_i\}\ .$$

The bias term (2.1) can be calculated once the knowledge of "relationship" between S_{n-1} and S_n is provided and this is possible only if the nature of the problems is specifically given. In this case, as we shall see in the next section, some better estimators are always

available. Here, "better" means smaller "biases." The key idea of constructing these better estimators is to estimate X_{n+1} by the current sample $\{X_1, \ldots, X_n\}$ first. The final estimate is obtained as if we had "$n+1$" observations. The idea is closely related to the idea of the EM algorithm (see Dempster et al (1977)).

3. Examples

3.1 Species Problems

In this section we shall continue our discussion of species problems introduced in Section 1. The problem of estimating the total probability of unseen species can be put in the framework of general coverage problem as in the previous section. Let E=$\{e_1, e_2, \ldots\}$ be the possible distinct species with probabilities $p_1, p_2, \ldots$, being selected in a single experiment. Let Δ denote the set of all positive integers. Let us make a natural correspondence between the outcomes space E and set Δ by "$e_i \leftrightarrow i$." The correspondence allows us to treat X_j as random variable such that $\{X_j = i\} \iff$ the j^{th} trial results an outcome e_i. It follows that in this case Ω=Δ,$F = 2^{\Delta}$ and $P\{X = i\}$=p_i for $i \in \Delta$.
Having observed $X_1, X_2, \ldots, X_n$, the collection of unseen species can be expressed as

$$S_n = S(X_1, X_2, \ldots, X_n; P) = \{j; j \notin \{X_1, \ldots, X_n\}\} \subset \Delta \, .$$

Let g denote an identity map from Ω to Ω, i.e., $g(i) = i$. The problem of estimating the total probability of unseen species is thus equivalent to estimating the probability of $g(X_{n+1}) \in S_n$ given S_n. More precisely,

$$P\{g(X_{n+1}) \in S_n | S_n\} \, .$$

According to the previous section, the $(n-1)$-estimate as in (i') is

$$\frac{1}{n} \sum_{j=1}^{n} I_{S_{n-1,j}}(X_j) = \frac{n_1}{n} \, , \tag{3.1.1}$$

where

$$S_{n-1,j} = \bigcup_{i \neq j}^{n} \{X_i\} = A_{n,j} \, .$$

Suppose we want to estimate the total probability of all species that appear $r (r \geq 1)$ times in the sample. By a similar argument,

$$S_n(r) = S(X_1, X_2, \ldots, X_n; P, r)$$
$$= \{j; \sum_{i=1}^{n} I_{X_i}(j) = r, j \in \Delta\}$$

and

$$S_{n-1,j}(r) = S(A_{n,j}; P, r) = \{i; \sum_{X_{i'} \in A_{n,j}} I_{X_{i'}}(i) = r, i \in \Delta\} .$$

Since

$$X_j \in S_{n-1,j}(r) \iff X_j \in S_n(r+1) ,$$

it follows from this fact that the $(n-1)$-estimate in this case (as in (i') again) takes the form

$$\frac{1}{n} \sum_{j=1}^{n} I_{S_{n-1,j}(r)}(X_j) = \frac{(r+1)n_{r+1}}{n} \tag{3.1.2}$$

which is formula (1.8).

The "Biases" of $(n-1)$-estimates. From (2.1) and (3.1.1), the "bias" of $(n-1)$-estimate in estimating $P\{X_{n+1} \in S_n | S_n\}$ is

$$\mathrm{E}\left\{\frac{n_1}{n} - \frac{n_1 + \delta}{n+1}\right\} ,$$

$$\text{where } \delta = \begin{cases} 1 & \text{if } X_{n+1} \notin \{X_1, \ldots, X_n\} \\ 0 & \text{if } X_{n+1} \text{ occurred at least twice among } \{X_1, X_2, \ldots, X_n\} \\ -1 & \text{if } X_{n+1} \text{ occurred once among } \{X_1, \ldots, X_n\}. \end{cases}$$

It follows trivially that

$$|\text{ bias of } (n-1)\text{–estimate}| \leq \frac{2}{n+1} = 0(\frac{1}{n}) . \tag{3.1.3}$$

The knowledge between the relationship of S_{n-1} to S_n enables us to construct a "better" estimate of which the bias is of order $(\frac{1}{n^2})$ contrast with the order of $(\frac{1}{n})$ provided by the previous $(n-1)$-estimate. The construction can be described heuristically as follows.

Let n_1' denote the number of species appearing once in the sample $\{X_1, X_2, \ldots, X_n, X_{n+1}\}$. Since X_{n+1} is missing, we cannot observe n_1', but instead we can estimate n_1', based on $\{X_1, X_2, \ldots, X_n\}$. Let $\hat{n}_1'$ denote this estimate which is defined by

$$\begin{cases} \hat{n}_1' = n_1 + 1 & \text{with prob. } \frac{n_1}{n} \\ \hat{n}_1' = n_1 & \text{with prob. } (1 - \frac{n_1}{n} - \frac{2n_2}{n}) \\ \hat{n}_1 = n_1 - 1 & \text{with prob. } \frac{2n_2}{n}. \end{cases}$$

The expected value of $\hat{n}_1'$ given $(n_1, n_2, \ldots)$ is

$$\text{(3.1.4)} \qquad \mathrm{E}\left(\hat{n}_1' | (n_1, n_2, \ldots,)\right) = n_1 + n_1 n^{-1} - 2n_2 n^{-1} \ .$$

The final estimate of estimating the total probability of unseen species in the sample $\{X_1, \ldots, X_n\}$ is

$$\text{(3.1.5)} \qquad \frac{\mathrm{E}(\hat{n}_1' | (n_1, \ldots,))}{n+1} = \frac{(n_1 + n_1 n^{-1} - 2n_2 n^{-1})}{n+1} \ .$$

The fact that the bias of this estimate is of order $0(\frac{1}{n^2})$ can be seen by noting that

$$\text{(3.1.6)} \qquad \mathrm{E}(\frac{n_1}{n} - \frac{2n_2}{n} - \delta) = \mathrm{E}(\frac{n_1}{n} - \frac{2n_2}{n} - \frac{n_1'}{n+1} + \frac{2n_2'}{n+1}) \ ,$$

where n_2' is the number of species appearing twice among $\{X_1, X_2, \ldots, X_n, X_{n+1}\}$.
It is clear that $|n_1 - n_1'| \leq 1$ and $|2n_2 - 2n_2'| \leq 2$ with probability one. It follows from (3.1.5) and (3.1.6) that the absolute bias of (3.1.5) is bounded by $\frac{3}{n(n+1)}$, which is of order $(\frac{1}{n^2})$.

One can mimic the above idea to find an estimator which is "better" than (3.1.2) in estimating the total probability of all species tht appear $r (r \geq 1)$ times in the sample. The improved estimator is

$$\text{(3.1.7)} \qquad [(r+1)n_{r+1} + n_{r+1}\Big((r+1)n_{r+1} - (r+2)n_{r+2}\Big)n^{-1}](n+1)^{-1} \ ,$$

which has smaller bias.

3.2 Estimating the Volume of a Convex Set in $\Re^k$.

The problem of estimating the volume of a certain convex set can be described as follows:

Let V denote a certain unknown convex set with finite volume in $\Re^k$. The data in this problem consists of independent random samples $X_1, X_2, \ldots, X_n$ uniformly distributed over V. The first question we want to ask is: having observed $X_1, X_2, \ldots, X_n$, how do we estimate vol (V)?

To answer this question, we first write down the joint likelihood of $X_1, \ldots, X_n$ as

$$\text{(3.2.1)} \qquad \text{Lik}\,(X_1, X_2, \ldots, X_n | V) = \Big[\frac{1}{\text{vol}(V)}\Big]^n \prod_{i=1}^n I_V(X_i)$$
$$= \Big[\frac{1}{\text{vol}(V)}\Big]^n I(V_n \subset V)\,.$$

where $V_n = V_n(X_1, X_2, \ldots, X_n)$ is the convex hull formed by $\{X_1, X_2, \ldots, X_n\}$, and $I(A \subset B) = 1$ if $A \subset B$, 0 otherwise.

It is easy to see from (3.2.1) that V_n, the convex hull formed by $\{X_1, X_2, \ldots, X_n\}$, is a sufficient statistic of V, according to Neyman's factorization theorem. This suggests that a reasonable estimate of vol(V) should be a function of V_n, the sufficient statistic of V.

To construct an estimate of vol(V), we first consider the problem of estimating the conditional probability $P(X_{n+1} \in V_n | V_n)$. As we shall see below, this problem can be treated as a special case of our general coverage problem.

Let $\Omega = V = \Delta$, and let F be the usual Borel field on V. Let P be the probability measure uniformly distributed over V. Define $g(w) = w$, the identity map from V to V. If we define $S_n = S(X_1, \ldots, X_n; P) = V_n(X_1, X_2, \ldots, X_n)$, the $(n-1)$-estimate of $P(X_{n+1} \in V_n | V_n)$ is

$$\text{(3.2.2)} \qquad \frac{1}{n} \sum_{j=1}^n I_{V_{n-1,j}}(X_j)\,,$$

where $V_{n-1,j}$ is the convex hull formed by $\bigcup_{i \neq j}^n \{X_i\}$. Since

$$P(X_{n+1} \in V_n | V_n) = \int_{V_n} \Big(\frac{1}{\text{vol}(V)}\Big) dw = \frac{\text{vol}(V_n)}{\text{vol}(V)}\,,$$

it follows that

$$\text{(3.2.3)} \qquad \text{vol}(V) = \frac{\text{vol}(V_n)}{P(X_{n+1} \in V_n | V_n)}\,.$$

Substitute $P(X_{n+1} \in V_n|V_n)$ by (3.2.2), the $(n-1)$-estimate of vol(V) is

$$\hat{\text{vol}}_{n-1}(V) = \text{vol}(V_n) \cdot \left[\frac{1}{n}\sum_{j=1}^{n} I_{V_{n-1,j}}(X_j)\right]^{-1} . \tag{3.2.4}$$

Like Section (3.1), the estimates (3.2.2) and (3.2.4) can be further improved. From (3.2.2), the $(n-1)$-estimate of $P(X_{n+1} \notin V_n|V_n)$ is

$$\frac{1}{n}\sum_{j=1}^{n}[1 - I_{V_{n-1,j}}(X_j)] = \frac{\#\text{ of vertices of } V_n}{n} . \tag{3.2.5}$$

Let vtx (U) denote the set of vertices of a convex polyhedron U in $\Re^k$, applying the similar idea of (3.1.4)-(3.1.7) to the current situation, we end up with a modified estimate (of $P(X_{n+1} \notin V_n|V_n)$)

$$\frac{\#\{\text{vtx}(V_n)\} + \frac{1}{n}\sum_{j=1}^{n}[\#\{\text{vtx}(V_n)\} - \#\{\text{vtx}(V_{n-1,j})\}]}{n+1} , \tag{3.2.6}$$

where # {vtx (U)}= number of vtx(U) for a convex polyhedron U. The modified estimates of $P(X_{n+1} \in V_n|V_n)$ and vol(V) are thus

$$1 - \{\#\{\text{vtx}(V_n)\} + \frac{1}{n}\sum_{j=1}^{n}[\#\{\text{vtx}(V_n)\} - \#\{\text{vtx}(V_{n-1,j})\}]\}(n+1)^{-1} \tag{3.2.7}$$

and

$$\text{vol}(V_n) \cdot \{1 - [\#\{\text{vtx}(V_n)\} + \frac{1}{n}\sum_{j=1}^{n}[\#\{\text{vtx}(V_n)\} - \#\{\text{vtx}(V_{n-1,j})\}]](n+1)^{-1}\}^{-1} , \tag{3.2.8}$$

respectively.

It is not difficult to check that the "biases" of estimates (3.2.6) and (3.2.7) are of smaller order ($0(\frac{1}{n^2})$, in fact) than those of (n-1)-estimates provided by (3.2.5) and (3.2.2). Since the arguments to verify this fact are very similar to those given in Section 3.1, we omit it.

The problem of estimating the volume of newly covered area if an additional observation X_{n+1} is taken can thus be estimated by the (n-1)-estimate.

$$\frac{1}{n}\sum_{j=1}^{n} \text{vol}[V_n \backslash V_{n-1,j}] = \frac{\text{vol}(\Delta_n)}{n} \quad \text{(say)} \tag{3.2.9}$$

where

$$\Delta_n = \bigcup_{j=1}^{n} \{V_n \backslash V_{n-1,j}\} .$$

A "better" estimate, using the similar idea of (3.1.4)-(3.1.7) again, can be expressed as

$$\frac{\text{vol}(\Delta_n) + \frac{1}{n}\sum_{j=1}^{n} [\text{vol}(\Delta_n) - \text{vol}(\Delta_{n-1,j})]}{n+1} , \tag{3.2.9'}$$

where

$$\text{vol}(\Delta_{n-1,j}) = \sum_{j' \neq j} \text{vol}[V_{n-1,j} \backslash V_{n-2,jj'}] ,$$

and $V_{n-2,jj'}$, is the convex hull formed by $\bigcup_{\substack{h \neq j \\ h \neq j'}} \{X_h\}$.

Before we move on to the next application, let us consider a simple example which may add some heuristic feeling to what we have done so far.

Example 3.1. Suppose $X_1, X_2, \ldots, X_n$ are *iid* from $U(\theta_1, \theta_2)$, with unknown parameter θ_1 and θ_2. The "volume" (length, in fact) of the current convex set is $\theta_2 - \theta_1$. Let $X_{(1)} < X_{(2)} \ldots, < X_{(n)}$ be the ordered values of $\{X_i\}_{i=1}^n$. It follows from previous discussion, the (n-1)-estimate of $P\,(X_{n+1} \in (X_{(1)}, X_{(n)}) | (X_{(1)}, X_{(n)})$ is

$$\frac{1}{n}\sum_{j=1}^{n} I_{V_{n-1,j}}(X_j) = \frac{n-2}{n} . \tag{3.2.10}$$

In fact, from (i) of Section 2 this (n-1)-estimate is an unbiased estimate of $P(X_n \in (X_{(1)}, X_{(n-1)}))$ based on n-1 observations $\{X_i\}_{i=1}^{n-1}$. The "better" estimates of $P(X_{n+1} \in (X_{(1)}, X_{(n)}) | (X_{(1)}, X_n))$ and $\theta_2 - \theta_1$, are (from (3.2.7)) thus

(3.2.11) $$1-\frac{2}{n+1}=\frac{n-1}{n+1}$$

and

(3.2.12) $$(X_{(n)}-X_{(1)})\frac{n+1}{n-1}\,,$$

respectively.

It is heuristically clear that the volume of V_n would tend to the volume of V as n goes to infinity. It is desired to find the rate (and distribution, if possible) that how fast the volume of V_n tends to that of V as n becomes large. As an application, we shall show in the following that the problem can be solved in $\Re^2$ via the interpretation together with a recent result of Groeneboom (1988). Let N_n be the number of vertices of V_n. If V is a convex polygon in $\Re^2$ with r edges, it was shown in Rényi and Sulanke (1963) that

$$EN_n \sim \frac{2}{3}r\log n \quad \text{as } n\to\infty\,.$$

It was also shown in the same paper that $\frac{EN_n}{n^{\frac{1}{3}}}\to$ constant if V has a smooth boundary in $\Re^2$. Since then much work has been done in this direction: Efron (1965), Geffroy (1959, 1961), Raynaud (1970), Eddy and Gale (1981), Buchta (1984), and Schneider (1987) among others.

In his recent paper, Groeneboom (1988) obtained some interesting results which will be stated as a proposition.

Proposition 3.1. (Groeneboom (1988))

(1) If V is a convex polygon with r vertices, then, as $n\to\infty$,

$$(N_n-\frac{2}{3}r\log n)/\sqrt{\frac{10}{27}r\log n}\xrightarrow{\mathcal{L}}N(0,1)$$

(2) If V is the unit disk on the plane, then, as $n\to\infty$,

$$(N_n-2\pi C_1 n^{\frac{1}{3}})/\sqrt{2\pi C_2 n^{1/3}}\xrightarrow{\mathcal{L}}N(0,1)\,,$$

where C_1, C_2 are two positive constants between zero and one.

From (3.2.5), the (n-1)-estimate $\frac{N_n}{n}$ is an unbiased estimate of $P(X_n\notin V_{n-1})=1-\frac{E(\mathrm{vol}(V_{n-1}))}{\mathrm{vol}(V)}$, that is,

$$\mathrm{E}(N_n) = \frac{n[\mathrm{vol}(V) - \mathrm{E}(\mathrm{vol}(V_{n-1}))]}{\mathrm{vol}(V)} .$$

It follows that

$$\mathrm{E}(N_n) = \frac{n[\mathrm{vol}(V) - \mathrm{E}(\mathrm{vol}(V_{n-1}))]}{\mathrm{vol}(V)} \tag{3.2.13}$$

$$= \begin{cases} \frac{2}{3} r \log n + o((\log n)^{+1/2}) & \text{if V is a polygon with r vertices} \\ 2\pi C_1 n^{\frac{1}{3}} + o(n^{+1/6}) & \text{if V is the unit disk.} \end{cases}$$

Combining (3.2.13), Proposition 3.1, and the fact that $\frac{N_n}{n}$ is an unbiased estimate of $P(X_n \notin V_{n-1})$, we have proved the following result.

Theorem 3.1

(1) If V is a convex polygon with r ($r \geq 3$) vertices, then, as $n \to \infty$,

$$n\left[\frac{N_n}{n} - P(X_n \notin V_{n-1})\right] \Big/ \sqrt{\frac{10}{27} r \log n} \xrightarrow{\mathcal{L}} N(0,1) \tag{3.2.14}$$

$$n\left[\frac{N_n}{n} - P(X_{n+1} \notin V_n)\right] \Big/ \sqrt{\frac{10}{27} r \log n} \xrightarrow{\mathcal{L}} N(0,1) \tag{3.2.15}$$

$$n\left[\frac{N_n}{n}\,\mathrm{vol}(V_n) - \mathrm{E}[\mathrm{vol}(V \backslash V_{n-1})]\right] \Big/ \left[\sqrt{\frac{10}{27} r \log n} \cdot \mathrm{vol}(V)\right] \xrightarrow{\mathcal{L}} N(0,1) \tag{3.2.16}$$

and

(2) If V is the unit disk in the plane, then, as $n \to \infty$, we have

$$P(X_n \notin V_{n-1}) \approx 0(n^{\frac{-2}{3}})$$

and

$$n^{5/6}\left[\frac{N_n}{n} - P(X_n \notin V_{n-1})\right] \Big/ \sqrt{2\pi C_2} \xrightarrow{\mathcal{L}} N(0,1) \tag{3.2.17}$$

$$(3.2.18) \qquad n^{5/6}\{\frac{N_n}{n}\,\mathrm{vol}(V_n) - \mathrm{E}[\mathrm{vol}(V\backslash V_n)]\}/\sqrt{2\pi C_2}\,\mathrm{vol}(V) \xrightarrow{\mathcal{L}} N(0,1)\ .$$

Note that (3.2.16) follows from the fact that

$$\mathrm{E}(\mathrm{vol}(V\backslash V_{n-1})) = \mathrm{vol}(V)P(X_n \notin V_{n-1})$$

and

$$\frac{N_n}{n}[\mathrm{vol}(V_n) - \mathrm{vol}(V)] = o_p(\frac{1}{n})\ ,$$

since

$$\frac{N_n}{n} \approx 0_p(\frac{\log n}{n}) \Rightarrow 1 - \frac{\mathrm{vol}(V_{n-1})}{\mathrm{vol}(V)} = 0_p(\frac{\log n}{n}) \Rightarrow \mathrm{vol}(V_n) - \mathrm{vol}(V) = 0_p(\frac{\log n}{n})\ .$$

Remark. In the case that V is a general convex set with smooth boundary, the results in (2) still hold, but with C_2 replaced by

$$C_2' = C_2(\pi/\mathrm{vol}(V))^{1/3}\int_{\partial V} k(s)^{1/3}ds/2\pi\ ,$$

where ∂V is the boundary of V, $k(s)$ is the curvature function of arc length. For detail, see Rényi and Sulanke (1963), and Groeneboom (1988).

Some implications deserve further discussion here. From (3.2.5), the probability of new observation X_{n+1} will fall outside the convex hull formed by the sample $\{X_1,\ldots,X_n\}$ is determined by the knowledge about the number of vertices of the conves hull. This result (i.e., (3.2.5)) holds for any distribution on $\Re^k$ and any $k \geq 1$. However, to estimate the volume of a convex body, the uniform distribution is used to create the relation like (3.2.3).

We don't have a general theorem like Theorem 3.1 in $\Re^k$ when $k \geq 3$ simply because a more general version of Proposition 3.1 is not available at the moment. However, from an applied point of view, we can always estimate the volume of a convex figure by Formula (3.2.4), and the vertices of V_n will provide us with information about $V \setminus V_n$. It seems to this author that almost all relevant information about $V \setminus V_n$ is within the set of vertices of V_n. This point will be further justified in Section 4 in terms of species problem.

The following problem is of interest:
Let V be a smooth convex figure. We know that $\frac{\text{vol}(V\setminus V_n)}{\text{vol}(V)} \approx 0(n^{-1})$ in $\Re^1$, and $0(n^{-\frac{2}{3}})$ in $\Re^2$ (from (2) of Theorem 3.1). What are the ratios in $\Re^k$ when $k \geq 3$. A less ambitious problem is to find the increasing rate $\frac{r_{k+1}}{r_k}$ of these ratios $\{r_1, r_2, \ldots\}$, where r_k stands for the ratio in $\Re^k$.

3.3 The Missile Problems

n missiles are delivered and landing at a certain target area which is usually much larger than the "effective area" caused by the explosion of a single missile. The "effective area" here can be referred to as a "covered area" in the present terminology. The problems we are interested in are: (1) if the $n+1^{\text{th}}$ missile is fired, what is the chance that this additional missile would involve area which was not covered previously? (2) How large is the newly covered area? (3) How many more missiles need to be fired in order to cover 90% of the target area?
To answer these types of questions, we introduce a simple model which seems to reflect the real situation reasonably close.

Let Δ denote the target area where the missiles would fall. Assuming that the locations of landing for all missiles are independent of each other and follow a certain unknown distribution G over Δ, let $Y_1, Y_2, \ldots, Y_n$ denote these n landing points. For each landing point Y_i, there is a covered area $B(Y_i, r_i)$ associated with Y_i, where $B(Y_i, r_i)$ denotes the intersection of Δ and the disk with center Y_i and random radius r_i. Note that each r_i may depend upon Y_i, but r_i and r_j are independent for different i,j since Y_i and Y_j are independent. If we let $X_i = B(Y_i, r_i)$ and $g(X_i) = Y_i$ for all $1 \leq i \leq n$, it is clear that the current model is within the framework of our general coverage problem described in Section 2.

The chance that the $(n+1)^{\text{th}}$ missile would land at "uncovered area" can be written as

$$(3.3.1) \qquad P(g(X_{n+1}) \notin S_n | S_n) \text{ , where } S_n = S_n(X_1, \ldots X_n; P) = \bigcup_{i=1}^{n} \{X_i\} .$$

From Section 2, the (n-1)-estimate is

$$(3.3.2) \qquad \frac{1}{n}\sum_{j=1}^{n}[1 - I_{S_{n-1,j}}(Y_j)] = \frac{\#\text{ of } \{Y_j; Y_j \notin \bigcup_{i\neq j}\{X_i\}\}}{n} .$$

where

$$S_{n-1,j} = \bigcup_{i \neq j} \{X_i\} .$$

Let us define $n_1(S_n)$=# of $\{Y_j; Y_j \notin S_{n-1,j}\}$ for brevity, and the (n-1)-estimate in (3.3.2) can thus be written as $\frac{n_1(S_n)}{n}$. Applying the similar idea of (3.1.4)-(3.1.7) to the current case, we come up with a "better estimate"

$$\frac{n_1(S_n) + \frac{1}{n}\sum_{j=1}^{n}[n_1(S_n) - n_1(S_{n-1,j})]}{n+1} . \tag{3.3.2'}$$

To estimate the size of newly covered area by the $(n+1)^{\text{th}}$ missile, it is easy to deduce from (ii) in Section 2 that the (n-1)-estimate is

$$\frac{1}{n}\sum_{j=1}^{n} \text{vol}[X_j \backslash S_{n-1,j}] = \frac{v_1(S_n)}{n} \quad \text{(say)} \tag{3.3.3}$$

where

$$v_1(S_n) = \sum_{j=1}^{n} \text{vol}[X_j \backslash S_{n-1,j}] .$$

Similarly, one can deduce a "better estimate" which is

$$\frac{v_1(S_n) + \frac{1}{n}\sum_{j=1}^{n}[v_1(S_n) - v_1(S_{n-1,j})]}{n+1} \tag{3.3.3'}$$

where

$$v_1(S_{n-1,j}) = \sum_{i \neq j}^{n} \text{vol}[X_i \backslash S_{n-2,ji}] \text{ for } 1 \leq j \leq n ,$$

and

$$S_{n-2,ji} = \bigcup_{\substack{h \neq i \\ h \neq j}} \{X_h\} .$$

4. Some limit theorems in species problem

In this section we shall present some large sample results for the various estimators derived from our interpretation in the species problem. The material of this section is somewhat technical. The idea used and the results obtained in this section are not limited to the species problem alone. With additional effort, it is expected to extend the idea to a more general situation which may cover all cases discusses in Section 3. However, in order to present the results simply and clearly we shall focus on the species problem.

Recall from Section 3.1 that $\{X_j = i\} \iff$ the j^{th} trial results on outcome $e_i \in \{e_1, e_2, \ldots\}$= outcome space. If for each outcome e_i there is a real value y_i (or a real vector y_i) associated with it, then we may ask the question: "Can one estimate the parameter associated with the unobserved species?" The general solution to this question will become apparent after we consider the following two simple examples.

Let $Y_j = y_i$ if $X_j = i$. The observed data are thus $\{(X_j, Y_j),\ 1 \le j \le n\}$. The outcome space is $\{(e_i, y_i)\}$.

Example 4.1. The mean.

In this case we are interested in the conditional mean of unobserved outcomes given $\{(X_i, Y_i)\}_{i=1}^n$, i.e.,

$$\int y dP(y|\mathbf{Y}_n)\ ,\ \text{where } \mathbf{Y}_n = (Y_1, Y_2, \ldots, Y_n)\ , \tag{4.1}$$

$$P(\mathrm{E}|\mathbf{Y}_n) = \sum_{y_j \in \mathrm{E}} p_j \varphi_j(0; n) \Big/ \sum_{j=1}^{\infty} p_j \varphi_j(0; n) \tag{4.2}$$

and E is any Borel set in $\Re$ (or in $\Re^k$ if y is a vector in $\Re^k$). The conditional distribution of $P(\mathrm{E}|\mathbf{Y}_n)$ can thus be written as

$$F(y|\mathbf{Y}_n) = P((-\infty, y]|\mathbf{Y}_n)$$

if $\{y_j\}$ are real-valued.

To estimate (4.1), we appeal to the interpretation. It is clear (from the interpretation) that the (n-1)-estimates of

$$\sum_{y_j \in \mathrm{E}} p_j \varphi_j(0; n) y_j \text{ and } \sum_{j=1}^{\infty} p_j \varphi_j(0; n)$$

are

$$\frac{1}{n}\sum_{\substack{j=1\\ Y_j\in E}}^{n} I_{S_{n-1,j}}(X_j)Y_j$$

and

$$\frac{1}{n}\sum_{j=1}^{n} I_{S_{n-1,j}}(X_j)=\frac{n_1}{n}\ ,$$

respectively. Recall that $S_{n-1,j}=\bigcup_{i\neq j}^{n}\{X_i\}=A_{nij}$. A natural $(n\text{-}1)$-estimate of $P(\mathrm{E}|\mathbf{Y}_n)$ is

$$\hat{P}(\mathrm{E}|\mathbf{Y}_n)=\Big[\frac{1}{n}\sum_{j=1}^{n} I_{S_{n-1,j}}(X_j)I(Y_j\in \mathrm{E})\Big]/(\frac{n_1}{n}) \tag{4.3}$$

$$=\Big[\sum_{j=1}^{n} I_{S_{n-1,j}}(X_j)I(Y_j\in \mathrm{E})\Big]/n_1\ .$$

The final $(n\text{-}1)$-estimate of conditional mean (4.1) is thus

$$\int y d\hat{P}(y|\mathbf{Y}_n)=\sum_{j=1}^{n} I_{S_{n-1,j}}(X_j)Y_j/n_1\ . \tag{4.4}$$

This simply tells us that, to estimate the conditional mean of unseen species one should use sample mean of the corresponding observations which occur only once in the sample.

Example 4.2. The median.

In this case we are interested in the median of $\{y_j; j\notin\{X_1,\ldots,X_n\}\}$. From the interpretation again, it is easy to check that the $(n\text{-}1)$-estimate is simply the sample median of Y_i of which the corresponding X_i occurs only once in the sample.

From these two examples it is not difficult to answer a more general question. If we are interested in a parameter $\theta=\theta(P(\cdot|\mathbf{Y}_n))$, which is a smooth function of $P(\cdot|\mathbf{Y}_n)$ as defined in (4.2), the naive $(n\text{-}1)$-estimate is thus $\hat{\theta}=\theta(\hat{P}(\cdot|\mathbf{Y}_n))$. Just how well is $\hat{\theta}$ as an estimate of θ? The success of estimating $\theta(P(\cdot|\mathbf{Y}_n))$ by $\theta(\hat{P}(\cdot|\mathbf{Y}_n))$ depends upon the magnitude of $\sum_{j=1}^{\infty} P_j\varphi_j(0;n)$, the total unobserved probability, which is estimated by $\frac{n_1}{n}$. The following propositions provide some theoretical justification of this estimate.

Proposition 4.1. Assuming that

$$EY^2 < \infty, n^{-\frac{1}{2}}\Big(\sum_{i=1}^{\infty} P_i(1-P_i)^{n-1}\Big)^{-1} = o(1)\ ,$$

and

$$\int y dP(y|\mathbf{Y}_n)$$

stays bounded in probability, then

$$\Big|\int y d\hat{P}(y|\mathbf{Y}_n) - \int y dP(y|\mathbf{Y}_n)\Big| \longrightarrow 0$$

in probability as $n \to \infty$.

Let

$$F_n(y) = F(y|\mathbf{Y}_n) = \sum_{y_j \le y} p_j\varphi_j(0;n) \Big/ \sum_{j=1}^{\infty} p_j\varphi_j(0;n)\ .$$

The estimate $\hat{F}_n(y) = \hat{F}(y|\mathbf{Y}_n)$ can be written as

$$\frac{1}{n}\sum_{i=1}^{\infty} \Psi_{i,n} \cdot I(y_i \le y) \Big/ \frac{1}{n}\sum_{i=1}^{\infty} \Psi_{i,n} = \frac{\#\text{ of } \{c_i; i \in A_1 \text{ and } y_i \le y\}}{n_1} \tag{4.5}$$

where

$$\Psi_{i,n} = \begin{cases} 1 & \text{if } i \text{ appears exactly once in } \{X_1, X_2, \ldots, X_n\} \\ 0 & \text{otherwise} \end{cases}$$

and

$$A_1 = \{X_i; \Psi_{X_i,n} = 1\}\ .$$

The following proposition shows that as an estimate of $F_n(\cdot)$ $\hat{F}_n(y)$ is uniformly consistent.

Proposition 4.2. Assuming that

$$n^{-\frac{1}{2}}\Big(\sum_{i=1}^{\infty} p_i(1-p_i)^{n-1}\Big)^{-1} = o(1)\ ,$$

then

$$\sup_y |\hat{F}_n(y) - F_n(y)| \longrightarrow 0$$

in probability.

We need some lemmas to prove these two propositions.

Lemma 1. Assuming that $k \geq 2$, then

$$\sum_{i=1}^{\infty} p_i^k (1-p_i)^n = 0(\frac{1}{n^{k-1}}) .$$

Proof. Since

$$(\frac{n-1}{k-1})^{k-1} \sum_i p_i^k (1-p_i)^n$$

$$= \sum_i \left[p_i^{\frac{k}{k-1}} (1-p_i)^{\frac{n}{k-1}} (\frac{n-1}{k-1}) \right]^{k-1}$$

$$\leq \sum_i \left[p_i^{\frac{1}{k-1}} (1-p_i)^{\frac{n}{k-1}} (1 + \frac{n-1}{k-1} p_i) \right]^{k-1} .$$

$$\leq \sum_i \left[p_i^{\frac{1}{k-1}} c^{-\frac{n}{k-1} p_i} \cdot c^{\frac{n-1}{k-1} p_i} \right]^{k-1} \quad \text{since } 1 - x \leq e^{-x}$$

$$= \sum_i \left[p_i^{\frac{1}{k-1}} c^{-\frac{p_i}{k-1}} \right]^{k-1}$$

$$= \sum_i p_i e^{-p_i} \leq 1 .$$

This completes the proof of Lemma 1.

Lemma 1'. If $E|Y| < \infty$ and $k \geq 2$, then

$$\sum_{i=1}^{\infty} p_i^k (1-p_i)^n y_i = 0(\frac{1}{n^{k-1}}) .$$

Proof. Since

$$|(\frac{n-1}{k-1})^{k-1} \sum_i p_i^k (1-p_i)^n y_i| \leq (\frac{n-1}{k-1})^{k-1} \sum_i p_i^k (1-p_i)^n |y_i| ,$$

the rest of the proof follows the same argument as that of Lemma 1.

Lemma 2. Assuming that $n^{-\frac{1}{2}} (\sum_i p_i (1-p_i)^n)^{-1} = o(1)$, then

(1) $$\left[\frac{n_1}{n} - \sum_i p_i(1-p_i)^n\right] / \sum_i p_i(1-p_i)^n = o_p(1)$$

(2) $$\left[\sum_i p_i\varphi_i(0;n) - \sum_i p_i(1-p_i)^n\right] / \sum_i p_i(1-p_i)^n = o_p(1)$$

Proof of (1). It suffices to show

(4.6) $$\mathrm{E}\left[\frac{n_1}{n} - \sum_i p_i(1-p_i)^n\right]^2 = 0(\frac{1}{n}) \ .$$

To see this, it is easy to check that under the assumption, the LHS of (4.6) is bounded by

(4.7) $$\sum_{i\neq j} p_ip_j(1-p_i-p_j)^{n-1} - \sum_{i\neq j} p_ip_j(1-p_i)^n(1-p_j)^n + 0(\frac{1}{n})$$

$$\leq \sum_{i\neq j} p_ip_j(1-p_i-p_j)^{n-1}(p_i+p_j) + 0(\frac{1}{n})$$

$$\leq 2\sum_i p_i^2 \sum_{j\neq i} p_j(1-p_i)^{n-1} + 0(\frac{1}{n})$$

$$\leq 2\sum_i p_i^2(1-p_i)^{n-1} + 0(\frac{1}{n}) = 0(\frac{1}{n}) \text{ (by Lemma 1)} \ .$$

This completes the proof of (1). Since the proof of (2) is quite similar, we omit it.

Lemma 3. Under the assumptions $n^{-\frac{1}{2}}(\sum_i p_i(1-p_i)^n)^{-1}=o(1)$ and $\mathrm{E}Y^2 < \infty$, we have

$$\mathrm{E}\left[\frac{1}{n}\sum_{j=1}^n I_{S_{n-1,j}}(X_j)(Y_j) - \sum_{i=1}^\infty p_i\varphi_i(0;n)y_i\right]^2 = 0(\frac{1}{n})$$

Proof. It is easy to see

$$\frac{1}{n}\sum_{j=1}^n I_{S_{n-1,j}}(X_j)(Y_j)$$

can be written as

$$\frac{1}{n}\sum_{i=1}^{\infty}\Psi_{i,n}y_i\ .$$

Now,

$$\begin{aligned}\mathrm{E}\{\frac{1}{n}\sum_i \Psi_{i,n}y_i\}^2 =&\frac{1}{n^2}\{\sum_i np_i(1-p_i)^{n-1}y_i^2\\ &+n(n-1)\sum_{i\neq j}p_ip_j(1-p_i-p_j)^{n-2}y_iy_j\}\\ =&\frac{1}{n}\sum_i p_i(1-p_i)^{n-1}y_i^2+\frac{n-1}{n}\sum_{i\neq j}p_ip_j(1-p_i-p_j)^{n-2}y_iy_j\end{aligned}$$

and

$$\begin{aligned}\mathrm{E}\Big(\sum_i p_i\varphi_i(0;n)y_i\Big)^2 =&\sum_i p_i^2(1-p_i)^n y_i^2\\ &+\sum_{i\neq j}p_ip_j(1-p_i-p_j)^n y_iy_j\ ,\end{aligned}$$

it follows that

$$n\ \mathrm{E}\left[\frac{1}{n}\sum_i\Psi_{i,n}y_i-\sum_i p_i\varphi_i(0;n)y_i\right]^2 \tag{4.8}$$

$$\begin{aligned}=&\sum_i p_i(1-p_i)^{n-1}y_i^2+(n-1)\sum_{i\neq j}p_ip_j(1-p_i-p_j)^{n-2}y_iy_j\\ &-2\sum_{i\neq j}np_ip_j(1-p_i-p_j)^{n-1}\\ &+n\sum_i p_i^2(1-p_i)^n y_i^2+n\sum_{i\neq j}p_ip_j(1-p_i-p_h)^n y_iy_j+0(\frac{1}{n})\ \text{(by Lemma 1')}\\ =&\sum_i p_i(1-p_i)^{n-1}[1+np_i(1-p_i)]y_i^2-\sum_{i\neq j}p_ip_j(1-p_i-p_j)^{n-2}y_iy_j+0(1)\end{aligned}$$

(This follows from Lemma 1, $\mathrm{E}|Y|^{r2}<\infty$, and similar argument in (4.7).)

$$\left(\begin{array}{l}\text{Note that } n\sum_{i\neq j} p_i p_j (1-p_i-p_j)^{n-1}(p_i-p_j)|y_i y_j| \\ \qquad \leq n\sum_i p_i^2(1-p_i)^{n-1}|y_i|\Big[\sum_j p_j|y_j|\Big] \\ \qquad \leq n\ \mathrm{E}\,|Y|\cdot 0(\frac{1}{n}) \text{ (by Lemma 1')} \\ \qquad =0(1)\end{array}\right)$$

It follows that (4.8) is

$$\leq \sum_i p_n y_i^2 + 0(1) < \infty\ .$$

This completes the proof of the lemma.

Proof of proposition 4.1

Rewrite

$$\int y d\hat{P}(y|\mathrm{Y}_n) - \int y dP(y|\mathrm{Y}_n)$$

as

$$D_n = \frac{b_n+\delta_n}{a_n+\epsilon_n} - \frac{b_n}{a_n}\ ; \tag{4.9}$$

where

$$a_n = \sum_i p_i\varphi_i(0;n),\ b_n = \sum_i p_i\varphi_i(o;n)y_i$$
$$\delta_n = \frac{1}{n}\sum_i \Psi_{i,n}y_i - b_n,\ \epsilon_n = \frac{n_1}{n} - a_n\ .$$

D_n can be further written as

$$\frac{a_n\delta_n - b_n\epsilon_n}{(a_n+\epsilon_n)a_n} = \frac{\delta_n}{a_n+\epsilon_n} - \frac{b_n\epsilon_n}{(a_n+\epsilon_n)a_n}\ . \tag{4.10}$$

By Lemma 2, $n^{-\frac{1}{2}}(\frac{n_1}{n})^{-1}=o_p(1)$. Since $\delta_n=o_p(\frac{1}{\sqrt{n}})$ by Lemma 3, and $\epsilon_n=0_p(a_n)$ by Lemma 2, it follows that

(4.11) $$\frac{\delta_n}{a_n + \epsilon_n} = o_p(1) ,$$

and

(4.12) $$\frac{\epsilon_n}{(a_n + \epsilon_n)} \cdot \frac{b_n}{a_n} = 0_p(1) \text{ .(since } b_n a_n^{-1} = o_p(1) \text{ by assumption)}$$

The proposition follows immediately from (4.11) and (4.12).

Proof of proposition 4.2

It is easy to check that

(4.13) $$\mathrm{E}\,(\frac{1}{n}\sum_i \Psi_{i,n} I(y_i \le y)) = \sum_i p_i(1-p_i)^{n-1} I(y_i \le y)$$

and

(4.14) $$\mathrm{E}\,(\sum_{y_i \le y} p_i \varphi_i(0;n)) = \sum_i p_i(1-p_i)^n I(y_i \le y) .$$

From Lemma 1, it is easy to see

$$\sup_y |\sum_i p_i(1-p_i)^{n-1} I(y_i \le y) - \sum_i p_i(1-p_i)^n I(y_i \le y)| = 0(\frac{1}{n}) .$$

Furthermore, with a similar argument as in Lemma 3, one can show that

(4.15) $$n\ \mathrm{E}\left[\frac{1}{n}\sum_i \Psi_{i,n} I(y_i \le y) - \sum_i p_i \varphi_i(0;n) I(y_i \le y)\right]^2 < M < \infty$$

for some positive M, independent of y. Proposition 4.2 is an immediate consequence of (4.12).

Acknowledgement. I wish to express my sincere thanks to Professor Herman Chernoff. The conversations I have had with him during this period of investigation were most valuable to me. I also wish to thank Professors Frederick Mosteller, Arthur Dempster, Donald Rubin, Persi Diaconis, and Arthur Cohen for the constructive comments they made during this study. Thanks also to Dr. B. H. Juang, from Bell Labs, who mentioned the language model to me and sent several useful references.

REFERENCES

Bahl, L.R., Jelinek, F., and Mercer, R.C. (1983) "Maximum likelihood approach to continuous speech recognition." *IEEE Trans. Pattern Analysis and Machine Intelligence* **5**, No. 2, 179-190.

Bickel, P.J. and Yahav, J.A. (1986) "On estimating the total probability of the unobserved outcomes of an experiment." *Adaptive Statistical Procedures and Related Topics.* IMS Lecture Notes. Edited by Van Ryzin. 332-337.

Buchta, C. (1984) "Stochastische approximation konvexer polygone." Z. Wahrscheinlichkeitstheor. *Verw. Geb* **67**, 283-304.

Clayton, M. and Frees, E. (1987) "Nonparametric estimation of the probability of discovering a new species." *JASA* **82**, 305-311.

Cohen, A. and Sackrowitz, H. (1987) "On estimating the probability of unobserved outcomes." Preprint.

Dempster, A., Laird, N., and Rubin, D. (1977) "Maximum likelihood from incomplete data vai the EM algorithm." (with discussion) *JRSS, B* **39**, No. 1, 1-38.

Diaconis, P. and Stein, C. (1983) "Decision theory." Lecture notes, Stanford University.

Eddy, W.F. and Gale, J.D. (1981) "The convex hull of a spherically symmetric sample." *Adv. Appl. Prob.* **13**, 751-763.

Efron, B. (1965) "The convex hull of a random set of points." *Biometrika* **52**, 331-343.

Efron, B. and Thisted, R. (1976) "Estimating the number of unseen species: how many words did Shakespeare know?" *Biometrika* **63**, No. 3, 435-447.

Estey, W.E. (1986) "The efficiency of Good's nonparametric coverage estimator." *Annals of Statistics* **14**, 1257-1260.

Geffroy, J. (1959) "Contribution à la théorie des valeurs extrémes." *Publ. Inst. Stat. Univ. Paris* **VIII**, 123-185.

Geffroy, J. (1961) "Localisation asymptotique du polyèdre d'appui dún échantillon Laplacien à k dimensions." *Publ. Inst. Stat. Univ. Paris* **X**, 212-228.

Good, I.J. (1953) "The population frequencies of species and the estimation of population parameters." *Biometrika* **40**, 237-264.

Good, I.J. and Toalmin, G. (1956) "The number of new species, and the increase in population coverage, when a sample is increased." *Biometrika* **43**, 45-63.

Groeneboom, P. (1988) "Limit theorems for convex hulls." *Probability Theory and Related Fields* **79**, 329-368.

Jelinek, F. (1976) "Continuous recognition by statistical method." *IEEE Proceedings* **64**, No. 4.

Katz, S.M. (1987) "Estimation of probability from sparse data for the language model component of speech recognizer." *IEEE Trans. Acoustics Speech and Signal Processing*, 400-401.

Raynaud, H. (1970) "Sur l'enveloppe convexe des nuages de points aléatoires dans R^n." *J. Appl. Probab.* **7**, 35-48.

Rényi, A. and Sulanke, R. (1963) "Über die konevxe Hülle von n zufällig gewählten punkten." *Z.W. Verw. Geb.* **2**, 75-84.

Robbins, H. (1956) "An empirical Bayes approach to statistics." *Proc. 3rd Berkeley Symp.* **1**, 137-163.

Robbins, H. (1968) "Estimating the total probability of the unobserved outcomes of an experiment." *Ann. Statist.* **39**, 256-257.

Robbins, H. (1977) "Prediction and estimation for the compound Poisson distribution." *Proc. Nat. Acad. Sci. USA* **74**, 2670-2671.

Schneider, R. (1987) "Random approximation of convex sets." Preprint. Mathematical Institute, Albertludwigs Univ., FRG.

Starr, N. (1979) "Linear estimation of the probability of discovering a new species." *Ann. Statist.* **1**, 644-652.